Objective Structured Clinical Examination Tests: Comparison with Traditional Clinical Examinations in Surgery

Daniel Ojuka, Nyaim Elly, Kiptoon Dan, Ndaguatha Peter

Department of Surgery, University of Nairobi

Correspondence to: Dr. Ojuka Daniel; Email: danielojuka@gmail.com.

Abstract

Background: Examination methods change over time, and audits are useful for quality assurance and improvement. **Objective:** Comparison of traditional clinical test and objective structured clinical examination (OSCE) in a department of surgery. **Methods:** Examination records of results of the fifth year MBChB examinations for 2012–2013 (traditional) and 2014–2015 (OSCE) were analyzed. Using 50% as the pre-agreed pass mark, the pass rate for the clinical examinations in each year was calculated and these figures were subjected to t-test to determine any significant differences in each year and in type of clinical test. P value of <0.05 determined significant statistical differences in the test score. **Results:** We analyzed 1178 results; most (55.6%) did OSCE. The average clinical scores examinations were 59.7% for traditional vs 60.1% for OSCE examination; basic surgical skills were positively skewed. **Conclusion:** OSCE in the same setting of teaching and examiners may give more marks than the traditional clinical examination, but it is better at detecting areas of inadequacies for emphasis in teaching.

Keywords: Clinical examination, Traditional, OSCE, Comparison

Introduction

Assessing clinical competence is one of the major tasks any medical teacher faces at the end of the term, and it gets gruesome at the end of the year. This is because the teacher's decision determines, in a short space of time, whether the candidate passes or fails on one hand, and on the other hand whether the safety of the community in which the candidate should be released is protected. This assessment, when incorporated within the course, provides relevant feedback to students and teachers in informing the students what is important and how they learn (1). It may inform teachers of areas to improve on to produce clinically competent doctors who will have better clinical performance in their internship years (2). The traditional examination tests a different competence; the student is given time to take history, perform physical examination and form an impression of the case (3). The candidate is then examined using an oral unstructured examination that tests the breadth and depth of the issue. The weakness of this method is its reliability. A good performance in one case does not predict a good performance in another, because of content. Students usually get one long case and a number of short cases, followed by an oral examination. One case was chosen because of the logistics. The implicit reason in choosing one case was, perhaps rather naïvely, the assumption that experienced doctors had the skills to immediately identify good or weak students on a single patient interaction, and that this was predictive of any patient interaction (4). It is not surprising therefore, that once the importance of context specificity was realized, both undergraduate and postgraduate clinical assessments have moved to the multi-station format of the OSCE (4). OSCE was introduced by Scottish doctor Ronald Harden in the 1970s, and has undergone variations depending on resources and context. It is the gold standard for medical examination because of its validity and reliability (5). We introduced OSCE in the Department of Surgery, University of Nairobi, Kenya, in 2014 for the undergraduate's surgical summative assessment for the final year MBChB candidates. In 2015, our second year, we analyzed the results to compare them with those of the years of traditional clinical exams, and to use the analysis to improve quality of training.

Methods

This was a retrospective analysis of examination results records of candidates from 2012–2015 at the Department of Surgery, University of Nairobi. Students undertake junior clerkship and senior clerkship. For the five-year program, the junior clerkship is 8 weeks long during the third year, with an end of rotation assessment that is mainly multiple-choice questions (MCQ). The fifth year has a 6-week rotation with MCQs, and a long case clinical examination at the end. The final assessment in surgery has the written component that has the traditional long essay and MCQ. Progressive assessment comprises marks obtained during the junior and senior clerkships while rotating in the various departments—ophthalmology, otorhinolaryngology, radiology, anesthesiology while considering attendance, log book and finally the clinical assessment during rotation. The final score in total is 700: the essay makes up 50, MCQ 150, progressive assessment 200, clinical examination 300. Emphasis is on clinical examination for a pass or fail decision. The clinical examination has been the traditional examination until 2014.

The traditional clinical examination consisted of one long case: the student was given 45 minutes to obtain history, perform physical examination, formulate diagnosis and differentials, and make notes on how they would manage the patient. The examination would take 15 minutes during which the candidate was given time to share the history and physical findings. A discussion then took place without a structured way of awarding marks. This was followed by four to five short cases, where the candidate was shown a patient with signs, or at times radiological films for quick diagnoses and discussion. The number of short cases given to the student depended on the performance of the student: the poor student would get more chances to prove themself. The candidate would then be taken through an oral examination where they were given equipment and other anesthesiology material and a discussion ensued after identifying or failing to recognize them. The student also went through the otorhinolaryngology and ophthalmology clinical stations, patterned in the manner of short cases. In this setting, the long case could be either orthopedics or general surgery. The short cases would be either a general surgical case or orthopedics or even specialty cases like burns, pediatric surgery, cardiothoracic cases or neurosurgical cases.

In 2014, when the OSCE was introduced, the department formed an OSCE committee that met and discussed and agreed that they would form 10 active stations and 2 rest stations. The 10 stations were: 4 stations for general Surgery (history taking, physical examination, management and communication skills); 3 orthopedic stations (history taking, physical examination and management); and 1 station each for anesthesiology, otorhinolaryngology and ophthalmology. The examination took 7 days. In 2015, the general surgery added one station (basic surgical skills) to make it five and replaced the communication station with interpretation of results. Due to the number of students, the number of resting stations varied between 5 and 7 each day for the 7 days. Each station took about 10 minutes, with one minute for transfer. The examination was performed in 3 ward settings, with each station asking the same question in each ward for the two years except in 2014, where the sites were 4 wards. History taking station had standardized patients with different cases each day, with emphasis on techniques. This was similar to the other stations. The questions were moderated by the OSCE committee and the marks were to be given as per a checklist.

Our examination takes about 3weeks. The essay paper has 6 questions: 2 are compulsory and the candidate must choose 2 from the other 4 questions. The structure is the 2 compulsory questions from general surgery and orthopedics, and the other 4arespread among the sub-specialties of pediatrics surgery, cardiothoracic surgery, plastic surgery and neurosurgery. The MCQs are 100 questions of best answer type. OSCE is as above.

We studied the records of the results for 5th year MBChB examinations for 2012, 2013, 2014 and 2015. Candidates within complete results were excluded. Marks scored in clinical examination, progressive assessments and the written paper were tabulated for each of the 4 years under study. Using 50% as the pre-agreed pass mark, the pass rate (percentage pass) for the clinical examinations in each year was calculated. We calculated the means and compared the mean for each type of examination, each year and each test. The mean scores (percent) for each type of examination were analyzed statistically using analysis of variance for two-factor. The two factors are the type of examination and the year of examination.

Results

In all, 1178 students completed their examination in the four years under analysis. The number of candidates seen increased progressively from 2012 to 2015—222, 301, 315, 340 respectively.

The mean score for various examination components also increased (Table 1).

Table 1: Descriptive statistics for the examinations across the years

Year	Measure	Progressive	Essay	MCQ	Clinical
2012	Mean	129.1	29.2	77.3	179.8
	SD	15.1	4.2	8.3	19.5
2013	Mean	135.5	28.1	77.9	178.3
	SD	11.3	2.8	9.5	27.6
2014	Mean	121.6	25.9	77.5	182.5
	SD	13.9	2.2	13.8	31.5
2015	Mean	133.7	29.3	91.1	190.1
	SD	15.5	3.3	8.4	26.3

Analysis of variance was performed for the examinations and Table 2 shows the results of that analysis.

Table 2: Analysis of variance for the various exams across the year

	Mean square	F-value
Progressive		
Between groups	12,000.086	61.022*
Within groups	196.651	
Essay		
Between groups	772.180	79.034*
Within groups	9.770	
MCQ		
Between groups	14,852.687	138.463*
Within groups	107.268	
Clinical		
Between groups	8,703.697	11.889*
Within group	732.083	

*All analyses were statistically significant (p<0.001)

The clinical examination reliability (Cronbach's α) from 2012 to 2015 was 0.58, 0.77, 0.79 and 0.72 respectively. An independent-samples *t*-test compared mean of traditional clinical examination score and OSCE clinical examination score. A significant difference was seen in the mean scores for traditional (Mean=178.96, SD=24.47, Variance=598.82) and OSCE (Mean=186.42, SD=29.18, Variance=851.33) clinical examinations; t (1176) =4.68, p=0.000. Pass rate for clinical examinations between 2012 and 2015 was 95%, 94%, 92.1% and 97.4% respectively. In general, the pass rate for the traditional examination in the two years was 94.5% while that for OSCE was 94.8%. From the six questions in essay, 2 were recall and 4 were comprehension questions. Most students picked the sixth question (224/340) because it was a recall question. The Pearson correlation between all types of examinations was weak (Table 3).

Table 3: Pearson *r*-correlation between the types of exams

	Progressive	Essay	MCQ	Clinical
Progressive	1	0.423	0.267	0.372
Essay	0.423	1	0.214	0.239
MCQ	0.267	0.214	1	0.232
Clinical	0.372	0.239	0.232	1

MCQs had only 13% problem-based questions, 87% were the recall type of questions. The Point Biserial ranged between –0.13and 0.36 with a Cronbach α of 0.53. In OSCE stations where students were required to demonstrate skills such as physical examination and basic surgical skills, there was positive skewedness compared with those talking skills such as history taking station or management stations (Figs. 1 and 2).
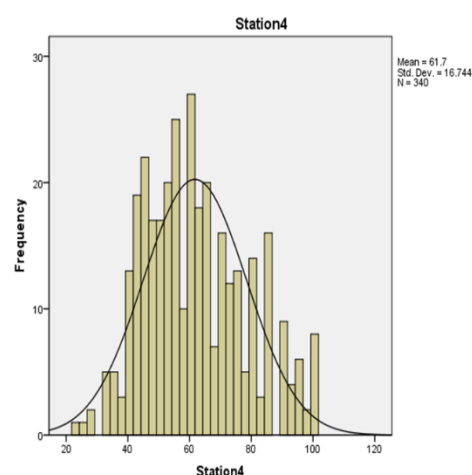

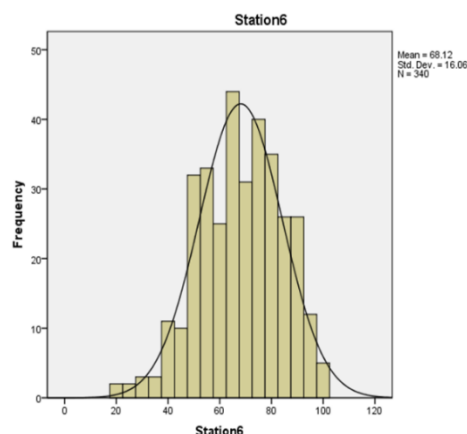Figure 1: Physical exam station

Figure 2: Management station

The standard deviation shows wide variation that may point to interobserver variability, given the stations were manned at different sites, testing the same questions but with different markers (Table 4).

Table 4: Station averages for the OSCE (2014 and 2015)

Year		St 1	St2	St 3	St 4	St 5	St 6	St 7	St 8	St 9	St 10
2015	Mean	65	60	66	62	65	68	64	58	58	77
	SD	18	17	13	17	17	16	14	16	16	14
2014	Mean	59	57	67	58	63	59	56	57	72	61
	SD	18	18	17	18	20	18	19	20	19	18

Discussion

The move towards objective assessment in medical education has seen the traditional methods of assessment of long case, short case and orals replaced with OSCE in most medical schools the world over (6). The replacement has been occasioned by poor reliability on long case and orals. Using the generalizability mathematical model, Wass et al. predicted that one student needs 10 long cases examined by two people to get a reliability of 0.80 (4). The reliability of orals varies between 0.50 and 0.80 (7), while that of ward rating is 0.25–0.37 (8). Our study reveals an average reliability of 0.68 (0.58, 0.77) for the traditional clinical examination and 0.76 (0.79, 0.72) for the OSCE clinical examination. These ratings are within the range of 0.46–0.88 that has been quoted in the literature for reliability of OSCE (9,10).

The mean score and pass rate for OSCE would suggest that either the OSCE was easy or candidates were of better quality in our study. This result is different from what is in the literature that suggests that OSCE was a downgrading score (11,12). However, considering the context, one would say that the reason could be because of examiners who might still be using the long case method. In that test there was a lot of prompting compared with the OSCE where the examiner ought to be an observer of the performance. Instead of just knowledge or "know how", OSCE is about "show how". When using a checklist, the OSCE examiner is a "recorder" of behavior rather than an "interpreter" of behavior (13,14). When one transitions from global rating to a checklist without clear learning on how to use the checklist, it may result in an upgrade when you give marks even after prompting the candidate.

The other reason why scores for OSCE may be higher in this study is because it has been shown that OSCE, because of the multi-station effects, evens out stringency whereas the traditional clinical examination has low chance of evening out stringency in awarding marks for examination (15). Another factor could be where standardized patients are used and not properly trained, the patient themself could help the candidate by giving cues (16).

When one considers the examination done by candidates in 2015 prototype of the examinations, a number of issues arise that need improvement for the quality of assessment to improve. The written essay questions' ability to test all what the students learn could be achieved by changing from the traditional essay to modified essay questions with use of clinical vignettes.

This will test higher levels of the Miller's pyramid (17). Though constructed response type of questions or their modification could test higher order thinking, they have been found to take time to construct and respond to. Their inter-rater reliability in marking is always low; hence some reviewers think they should not be used in any high-stake examination (18,19). The construction of our essays did not cover all the subjects, and some were context-free. These essays could not assess higher functions as they were meant to, hence the need to change. The Biserial point for most of the MCQ was low, with a very low number of problem-solving questions. These should be improved through training faculty in how to set quality multiple choice questions. Well-constructed selected–response questions with clinical vignettes have been shown to evaluate higher-order thinking in the modified Miller's pyramid (17). But our questions may need modification in that respect.

Our correlation test between MCQ and clinical examinations was very low, demonstrating that this examination tests different domains of knowledge or concepts (11,12). However, other studies found that they correlate very well (17). Low correlation points to poor construction of questions, as shown by other indices such as Point Biserial (17,20).

Assessment is one of the determinants of students' learning style (1). When students realize what is required is 'know-how' and not 'show-how', as is the traditional long case, they may resort to learning styles that are considered superficial as opposed to deep (17). This is because assessment is a high stakes examination where people fear failure and its consequences. In the traditional clinical examination where history and physical examination are not observed, it is easy for candidates to learn "know" and "know-how" and fail on the "show-how". That is seen in our OSCE results where the "show-how" stations scores are skewed to the left. If a deep learning style among students is desirable, we need to change the mode of assessment where what is valued is not the score of the "know" or "know-how" but the

progressive assessment is taken seriously and its marks used as a major determinant of a pass or a fail.

Conclusion

The OSCE in this study seems to upgrade students more than the former traditional long case clinical examination, though the examination in 2015 upgraded students in general. The study reveals a weakness in stations where "show-how" is required as opposed to "know-how".

References

1. Newble D, Jaeger K. The effect of assessments and examinations on the learning of medical students. Med Educ. 1983; 17(3):165–71.
2. Newble D. Assessing clinical competence at the undergraduate level. Med Educ. 1992; 26(6):503–11.
3. Ponnamperuma G, Karunathilake I, McAleer S, et al. The long case and its modifications: A literature review. Med Educ. 2009; 43(10):936–41.
4. Wass V, Van Der Vleuten C. The long case. Med Educ. 2004; 38(11):1176–80.
5. Hodges B. OSCE! Variations on a theme by Harden. Med Educ. 2003; 37(12):1134–40.
6. Vleuten C, Norman G, Graaff E. Pitfalls in the pursuit of objectivity: Issues of reliability. Med Educ. 1991; 25(2):110–8.
7. Muzzin L, Hart L. Oral examinations. In: Neufeld, Victor R, Norman G (eds), Assessing clinical competence. Springer, New York. 1985; 71–83.
8. Streiner D. Global rating scales. Assess Clin Competence N Y Springer. 1985; 119–41.
9. Walters K, Osborn D, Raven P. The development, validity and reliability of a multimodality objective structured clinical examination in psychiatry. Med Educ. 2005; 39(3):292–8.
10. Turner J, Dankoski M. Objective structured clinical exams: a critical review. Fam Med. 2008; 40(8):574–8.
11. Coovadia H, Moosa A. A comparison of traditional assessment with the objective structured clinical examination (OSCE). S Afr Med J. 1985; 67(20):810–2.
12. Harden R, Gleeson F. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ. 1979; 13(1):39–54.
13. Yudkowsky R, Park Y, Riddle J, et al. Clinically discriminating checklists versus thoroughness checklists: Improving the validity of performance test scores. Acad Med. 2014; 89(7):1057–62.
14. MacRae H, Vu N, Graham B, et al. Comparing checklists and databases with physicians' ratings as measures of students' history and physicalexamination skills. Acad Med. 1995; 70(4):313–7.
15. Finn Y, Cantillon P, Flaherty G. Exploration of a possible relationship between examiner stringency and personality factors in clinical assessments: a pilot study. BMC Med Educ. 2014; 14(1):1052.
16. Newble D, Hoare J, Sheldrake P. The selection and training of examiners for clinical examinations. Med Educ. 1980; 14(5):345–9.
17. Hift R. Should essays and other "open-ended"-type questions retain a place in written summative assessment in clinical medicine? BMC Med Educ. 2014; 14(1):249.
18. Palmer E, Duggan P, Devitt P, et al. The modified essay question: Its exit from the exit examination? Med Teach. 2010; 32(7):e300–7.
19. Palmer E, Devitt P. Assessment of higher order cognitive skills in undergraduate education: Modified essay or multiple-choice questions? Research paper. BMC Med Educ. 2007; 7(1):49.
20. Ramos P, Ramirez G, Vallejo E, et al. Performance of an objective structured clinical examination in a national certification process of trainees in rheumatology. Rheumatolclin. 2015; 11(4)215–20.